# Bibliometric Fusion:

## *An Open Science Collaborative Project on Research Collaboration Network Mapping*

***Shenmeng Xu**, Librarian for Scholarly Communications, Vanderbilt University Libraries, ORCID: 0000-0001-8475-0746 and **Steven J. Baskauf**, Data Science and Data Curation Specialist, Vanderbilt University Libraries, ORCID: 0000-0003-4365-3135*

### NUTRITION INFORMATION

This recipe is crafted for academic libraries interested in fostering collaborations with researchers, administrators, or students to leverage bibliometric data and methods for the purpose of building, analyzing, and understanding research collaborations, all while upholding a commitment to open science principles.

In a research collaboration network, nodes represent authors, and links among them signify coauthor relationships. Collaboration networks facilitate the identification of key contributors and influencers. Mapping collaboration networks aids in recognizing knowledge hubs and nodes that facilitate the sharing of expertise and resources. Understanding how collaborators are connected can reveal potential areas for innovation. When presented in an accessible manner, such as an interactive network map, this dish might provide insights for diverse stakeholders: prospective students seeking academic institutions, programs, and advisors; researchers seeking potential collaborators; administrators seeking insights into collaborations among institutions and individuals; and patients in search of expertise within specific physician subspecialties (see figure 1), among others.

In this collaboration, academic librarians from diverse subspecialties within information and library science serve as experts in bibliographic databases, scholarly communication, data handling, analytics, visualization, and data management. They provide crucial guidance and support for promoting open science practices. With their support, this collaboration model effectively addresses the challenges and resource limitations faced by researchers, administrators, and students in other fields when planning and implementing a project of this nature.

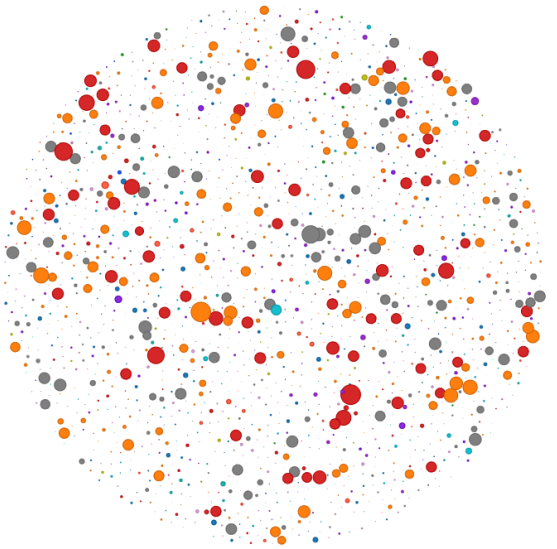

**Figure 1.** In this screenshot of a sample collaboration network depicting clinician researchers in a medical field in the United States, node size indicates Eigenvector centrality, and node color indicates community.

Drawing inspiration from the concept of fusion cuisine, this approach seamlessly integrates diverse data sources and tools, adapting them to meet customized analytical needs, constructing a holistic portrait of research collaboration patterns, incorporating domain-specific norms for comprehensive understanding, and presenting these findings in an innovative way.

### LEARNING OUTCOMES

After reading this recipe, cooks will be able to:

- identify the requisite resources and collaborators necessary for an open science project focused on mapping research collaboration networks;
- customize and apply the workflows to suit their specific requirements, enabling them to implement similar projects aimed at mapping collaboration networks; and
- evaluate and determine optimal approaches at each stage.

## NUMBER SERVED

Aligning with the principles of open science, this recipe is tailored for expansive tasting events within and beyond the academic community rather than private dinner parties.

## COOKING TIME

Depending on the project's scope and the level of collaboration among contributors, the preparation of this dish may span from one to several months.

## INGREDIENTS AND EQUIPMENT

The list below outlines the essential ingredients and equipment required for preparing this recipe, but adaptability and creativity are key. Cooks are encouraged to tailor their approach based on their unique goals and constraints.

- in-depth understanding of academic research and scholarly publishing
- knowledge of persistent identifiers (PIDs) referring to scholarly works, contributors, and organizations
- capability to query and collect data from bibliographic databases indexing scholarly publications
- permission to utilize the databases and data for research purposes
- skills and tools for effective data cleaning and wrangling
- proficiency in network analysis and visualization, supported by appropriate tools
- code and data sharing platforms
- reliable web hosting infrastructure

## PREPARATION

1. **Delineate project scope and research aims.** To determine the project scope, it is critical to identify the authors of interest. Depending on whether or not your project aims to focus on a definitive list of authors, the workflow differs (see figure 2).

One workflow involves starting with a definitive list of authors and then retrieving publication information. Author lists might be generated from specific institutions, departments, scholarly societies, laureates of awards, etc. Alternatively, cooks can start with a list of publications associated with a topic of interest, journal, or a conference and then extracting author information. The Cooking Method section will illustrate the former approach to creating an interactive collaboration network of researchers based on Scopus data.

2. **Establish collaboration and define team roles.** For this fusion dish, it is critical to have a team of cooks with diverse backgrounds. A possible team combina-

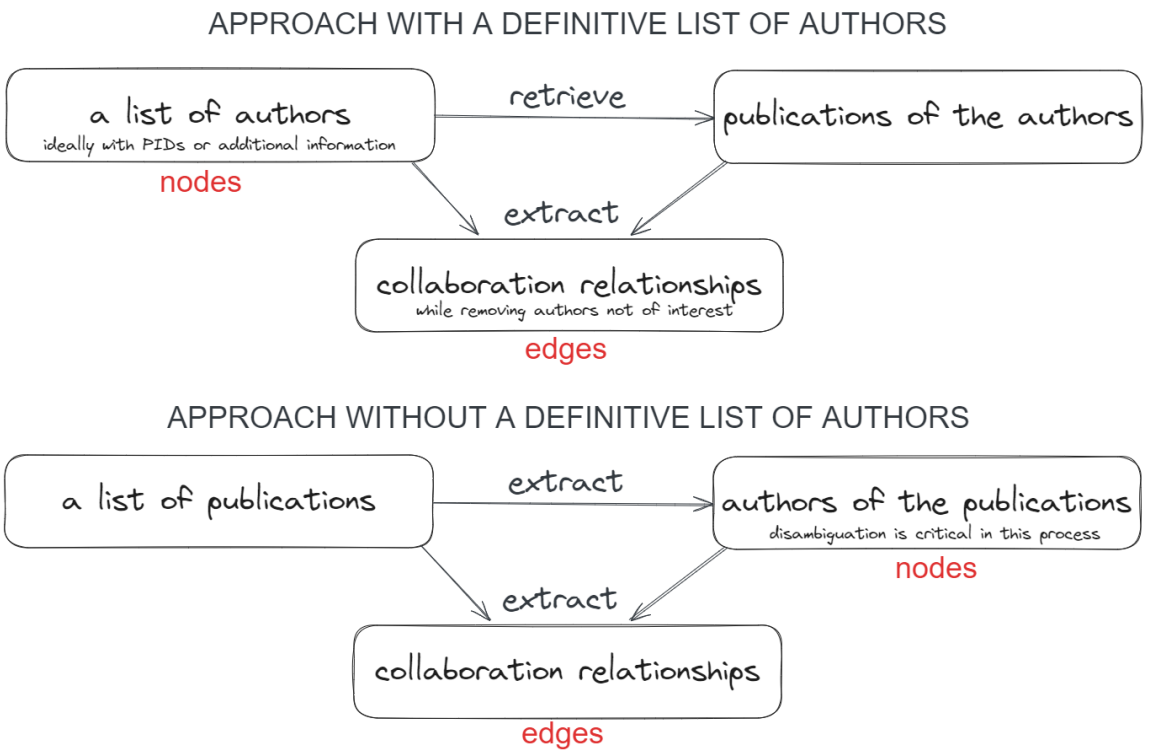


**Figure 2.** Two workflows depending on the availability of a definitive list of authors.

tion includes a scholarly communication librarian, a data science librarian, a subject librarian, a researcher, and a research assistant. It is recommended to have at least one domain expert that understands discipline-specific scholarly norms, such as common publication types, venues, collaboration and contributorship practices. Cooks will serve different roles. Some cooks might serve as consultants.

3. **Review the ingredients and equipment guideline above, and identify what best serves your project scope and research aims.** Following are the ingredients and equipment for the approach outlined in this recipe.
   - Databases: ROR, Scopus (including the Scopus Search API and the Author Search API)
   - PIDs: ROR ID, Scopus EID, DOI, PubMed PMID, Scopus Author ID
   - Jupyter Notebook and Python:
     - libraries for data collection: requests, json, pandas, NumPy, FuzzyWuzzy
     - libraries for network construction, analysis, and visualization: NetworkX, Pyvis, Matplotlib
   - Github

If you are missing any ingredients or equipment, refer to the Additional Resources section for preparation guidance.

## COOKING METHOD

1. **Determine sampling and data collection strategies.** Consider criteria such as field, affiliation, time window, geographic location, etc.

2. **Data collection, cleaning, and processing.** The goal is to prepare a node list (authors) and an edge list (collaboration relationships). Cooks are encouraged to tailor these steps based on their equipment and ingredients, ideally automating this process as much as possible.

Creating a node list: As illustrated in figure 1, when retrieving publications for a list of authors, the ideal scenario involves having PIDs for all authors. However, this is often not the case in reality. Some other information might help. For instance, if affiliation information is available for the authors, use their institutions' ROR ID as an additional restriction when matching author names to Scopus Author IDs. This will help eliminate the majority of false positives (authors with the same names at other institutions). Due to the potential lack of uniqueness in names even within institutions, manual review remains necessary. If PIDs are available, skip this step and use the PIDs to collect additional data of interest, such as citation counts, h-index, etc.

Creating an edge list*:* Query the Scopus Search API and retrieve metadata for publications associated with the Scopus Author IDs collected earlier. For each publication, you will need, at minimum, the publication's PIDs (Scopus EID, DOI, and PubMed PMID) as aids in subsequent cleaning and de-duplicating processes. Additionally, collect the Scopus Author IDs of all authors for these publications. Compare these Author IDs with those in the predefined list and exclude authors not of interest. Iterate through the list of publications and extract collaboration relationships and frequencies for all authors. This list of collaboration relationships constitutes the edge list.

Note: plan data collection accordingly depending on the API rate limits.

3. **Network construction**. Take the node and edge lists and construct the network using the NetworkX library. Cooks who use Python can refer to the Jupyter Notebook (see Additional Resources section below) for step-by-step guidance.

4. **Data analysis and visualization.** Use NetworkX to calculate network statistics.
   - Degree centrality indicates the number of collaborative relationships an author has with others.
   - Betweenness centrality measures how effectively an author serves as a bridging role and connects others who might not otherwise be linked.
   - Eigenvector centrality rises when an author collaborates with well-connected collaborators.
   - Closeness centrality indicates how quickly an author can reach everyone else in the network.

Utilize hierarchical clustering techniques, such as the Girvan–Newman or Louvain algo-

rithms, to group authors based on similarities or distances and detect communities in the network.

Next, use Pyvis to bring the network data to life by generating a dynamic and interactive web-based representation of the network (rendered using HTML and JavaScript). Customize the colors, shapes, sizes, and labels of the nodes and the edges to best present features of the network.

5. **Communication of findings.** Librarian cooks are encouraged to instruct in open science practices and actively promote open science principles throughout the cooking process. To enhance reproducibility, transparently share methodologies and data, leveraging platforms such as Github, OSF, Zenodo, Dryad, and your institutional repository for code and data dissemination. Additionally, hosting the interactive network on a web server will allow users to explore and interact with the network in an intuitive and engaging manner.

## CHEF'S NOTES

This recipe is highly customizable. Cooks are encouraged to tailor their approach based on their unique goals and constraints. For instance, consider using the search interface and batch download method instead of querying from APIs; opt for the free index OpenAlex rather than subscription-based tools; or choose user-friendly software that does not require coding, such as Gephi for network analysis, as an alternative to using Python.

Maintenance can be a challenge in collaborative projects like this. To make sure that the food has a long shelf life, it is critical to plan ahead and proactively implement a sustainable and comprehensive maintenance strategy.

## ADDITIONAL RESOURCES

Rieger, O. Y., & Schonfeld, R. C. (2023, April 24). Common scholarly communication infrastructure landscape review. *Ithaka S+R*. https://doi.org/10.18665/sr.318775

Menczer, F., Fortunato, S., & Davis, C. A. (2020). *A first course in network science*. Cambridge University Press. https://cambridgeuniversitypress.github.io/FirstCourseNetworkScience/

NetworkX, https://networkx.org/

Pyvis, https://pyvis.readthedocs.io/

Sample Jupyter Notebook for network analysis, https://github.com/ShenmengXu/acrl-os-cookbook/blob/main/Network%20Analysis%20Example%20for%20ACRL%20Open%20Science%20Cookbook.ipynb